Adam Pawłowski
(Uniwersytet Wrocławski)

KORPUSY CHRONOLOGICZNE I LEKSYKALNE SZEREGI CZASOWE JAKO NARZĘDZIA WYKRYWANIA SŁÓW KLUCZY I NEOSEMANTYZMÓW. KONCEPTUALIZACJA CZASU W KORPUSACH CHRONOLOGICZNYCH[1]

WSTĘP

Termin słowo klucz funkcjonuje dziś w informatologii, informatyce, medioznawstwie, literaturoznawstwie i lingwistyce, a ponadto – jako motywowana semantycznie, a przez to czytelna metafora – także w języku ogólnym. Jego kariera i niewątpliwa użyteczność związane są z rosnącą potrzebą syntetyzowania informacji rozproszonej w niemal bezkresnej przestrzeni Internetu – nie tylko w tekstach nowotworzonych, lecz także powstałych w minionych wiekach i dziś digitalizowanych. Dla informatologa słowo kluczowe[2], podobnie jak hasło przedmiotowe i deskryptor, jest fragmentem mikroopisu tekstu – na ogół, ale nie wyłącznie, naukowego lub medialnego – i służyć ma jego wyszukaniu, klasyfikowaniu i/lub pozycjonowaniu w systemie informacyjnym (por. Ścibor 1999; Bojar 2002; Babik 2010). Warto pamiętać, że syntetyzowanie treści dokumentu ma kilka poziomów: pierwszym jest tytuł[3], drugim abstrakt, streszczenie lub lid (charakteryzuje je struktura narracyjna), trzecim lista słów kluczy (deskryptorów, haseł przedmiotowych). Podstawowa różnica między słowami kluczowymi z jednej strony a hasłami przedmiotowymi i deskryptorami z drugiej jest taka, że te ostatnie dobierane są z tzw. słownika kontrolowanego danej dyscypliny, czyli zamkniętego

[1] Artykuł powstał w ramach konsorcjum CLARIN-PL, finansowanego ze środków MNiSW. Decyzja nr 3255/CLARIN ERIC/2015/0 z 16 lutego 2015 r.

[2] Zdaniem niektórych badaczy istnieje różnica między wyrażeniami „słowo klucz" i „słowo kluczowe": pierwsze należałoby do nauki o języku i dyscyplin pokrewnych, natomiast drugie do informatologii (por. Bojar 2002). Chociaż oba znaczenia są niemal synonimiczne (różni je nie tyle treść, co kontekst zastosowania i funkcja), dla zachowania spójności metodologicznej podział ten będzie w niniejszym artykule respektowany. W dalszej części artykułu będą stosowane zamiennie nazwy „słowa klucze" i „słowa tematyczne".

[3] Strategia tytułowania publikacji zmieniała się w ciągu stuleci. W zasadzie do drugiej połowy dziewiętnastego wieku można mówić o wielozdaniowych tytułach opisowych, pełniących funkcje dzisiejszego abstraktu i perytekstu (tytuł taki informował, streszczał, objaśniał i zachęcał do zakupu). W okresie późniejszym, przede wszystkim w wieku dwudziestym, tytuły zaczęto skracać, a funkcje informacyjne lub perswazyjne przejęły inne fragmenty publikacji (tylna okładka, skrzydełka oprawy, przedmowa), a częściowo także teksty istniejące poza jej fizyczną formą, lecz z nią powiązane (zapowiedzi, recenzje), określane jako epitekst (por. Genette 1987).

i ustalonego *a priori* zbioru terminów, podczas gdy repertuar potencjalnych słów kluczy nie jest ograniczony, przynajmniej pod względem semantycznym. Ponieważ katalogowane dokumenty są na ogół wielotematyczne, w praktyce pełne mikroopisy tworzone są przez zestawy kilku terminów, wskazujących na treści najważniejsze, najbardziej reprezentatywne, pozwalające na wyszukanie w systemie bazodanowym lub w sieci WWW (Babik 2010; Malak 2012).

Lingwistyczne podejście do procesu generowania i funkcji słów kluczy jest w ogólnych zarysach podobne do informatologicznego, aczkolwiek ma pewne cechy specyficzne. Owa specyfika wynika nie tyle z treści syntetyzowanych dokumentów, co z bardziej zróżnicowanych celów badań oraz właściwości przestrzeni cyfrowej, w której przebiega dziś większość komunikacji piśmienniczej. O ile perspektywa informatologiczna, jak wyżej wspomniano, jest zorientowana na charakterystykę konkretnych dokumentów i ma umożliwić ich wyszukanie lub pozycjonowanie (nawet w kolekcjach o objętości dziesiątek milionów pozycji), o tyle dla lingwistyki dokument jako jednostka badawcza nie jest szczególnie istotny. Celem lingwistów jest natomiast eksploracja danych tekstowych, dopuszczająca analizę dowolnych strumieni informacji zagregowanej, pochodzącej z różnych kanałów komunikacji. Słowa roku, miesiąca, tygodnia czy dnia – określane metaforycznie jako „słowa na czasie" – są właśnie efektem takiej eksploracji: ujawniają tematy dominujące w komunikacji w danym okresie, nie służą jednak tworzeniu charakterystyk meta-informacyjnych poszczególnych dokumentów. Tym, co łączy oba podejścia do problematyki słów kluczy (kluczowych, tematycznych), jest przekonanie, że wyrażają one w największym możliwym skrócie, a mimo to skutecznie i komunikatywnie, najważniejsze treści zawarte w danych tekstowych.

SŁOWA KLUCZE A „CZYTANIE WSPOMAGANE"

Rewolucja cyfrowa, poprzedzona skokowym wzrostem produkcji drukowanej w drugiej połowie dziewiętnastego wieku, doprowadziła do tego, że pełne skupienia, wnikliwe i pogłębione czytanie niewielkiej liczby tekstów, charakterystyczne dla kultury osiemnastego i dziewiętnastego wieku, stało się nieefektywne. Dotyczy to zresztą nie tylko tzw. czytelnika ogólnego. Nieznana wcześniej masowość produkcji tekstów i intensywność ich wymiany sprawiły, że nawet profesjonaliści utracili możliwość samodzielnego śledzenia piśmiennictwa w poszczególnych dyscyplinach nauki (por. Lem 1999; Cortada 2002, 2012; Castells 2007).

Przyswajanie nowej wiedzy, stanowiące podstawowy mechanizm adaptacji *homo sapiens* do zmieniającego się środowiska informacyjnego, może jednak dokonywać się z pomocą technologii automatycznego przetwarzania języka (*natural language processing* – NLP). Podobnie jak inne rozwiązania inżynieryjne, służące przekraczaniu granic fizycznych i zdolności kognitywnych człowieka, NLP umożliwia prowadzenie analiz wielkich zbiorów tekstów (analiza *big data*), wydobywanie z nich potrzebnych treści i ich klasyfikowanie.

Warto więc w tym miejscu postawić pytanie, czy syntetyzowanie informacji zawartej w wielkich masach tekstu za pomocą narzędzi NLP może być traktowane jako nowa forma lektury. Jeżeli czytanie zdefiniuje się jako akwizycję wiedzy przez interakcję człowieka i tekstu, aktywne współ-

uczestnictwo w tym procesie komputerów i oprogramowania, wykraczające poza funkcję nośnika, nie powinno być problematyczne. W literaturze przedmiotu istnieje zresztą pojęcie *distant reading* (por. Moretti 2013), które należałoby określić jako c z y t a n i e w s p o m a g a n e: polega ono właśnie na automatycznym wydobywaniu potrzebnej czytelnikowi informacji. Jego efektem mogą być streszczenia, infografiki, mapy geolingwistyczne, a także listy słów kluczowych lub tematycznych.

Automatyczną ekscerpcję „słów na czasie" można więc rozumieć jako jedną z metod czytania wspomaganego komputerowo. Polegałoby ono na monitorowaniu strumienia danych tekstowych w wybranych kanałach komunikacyjnych i wydobywaniu z nich tych leksemów, które – według przyjętych założeń – można uznać za reprezentatywne dla szczególnie znaczących zjawisk i procesów zachodzących w świecie pozajęzykowym. Badany obieg komunikacyjny można definiować, stosując dowolne kryteria, na przykład gatunkowe (teksty medialne, literackie, użytkowe, blogi, zapisy komunikatorów), stylistyczne (proza, poezja, dialog, tekst opisowy), socjologiczne (język grup wiekowych, zawodowych, społecznych) lub chronologiczne (teksty określonej epoki).

Można w tym kontekście przypomnieć, że czynność lektury przeszła w ciągu wieków liczne przeobrażenia (por. Ong 1992; Manguel 1996; Chymkowski 2004; Havelock 2006; Vandendorpe 2008) i opisana tutaj zmiana nie jest niczym niezwykłym – wpisuje się w ciąg efektów ubocznych towarzyszących tzw. rewolucjom medialnym lub przemianom technologicznym (por. Góralska 2012). Wynalazek pisma i jego upowszechnienie w starożytności lub (stosowanie do regionu) we wczesnym średniowieczu zepchnęły na margines kulturę oralną, wymagającą doskonałych zdolności pamięciowych. W Europie wczesnośredniowiecznej nowością było „czytanie oczami" – ciche, zindywidualizowane i szybsze od głośnej, publicznej lektury. Znany jest fragment *Wyznań* św. Augustyna, w którym autor pisze o tej szczególnej umiejętności, jaką posiadał Ambroży, biskup Mediolanu: „Gdy czytał, oczy przebiegały stronice, a umysł rozważał treść tekstu, język zaś był bezczynny i żaden dźwięk nie dobywał się z ust" (Augustyn 1987: 141). Jeśli autor tych słów uznał za godne uwagi podjęcie tematu cichej lektury w swoim dziele, musiał mieć ku temu podwód: było to zapewne zjawisko rzadkie, wyjątkowe, dla niektórych wręcz szokujące. We wczesnej starożytności ten sposób lektury w zasadzie nie był znany: „dziś bowiem właściwie bez poważnych zastrzeżeń przyjmuje się, że starożytni Grecy nie znali pojęcia czytania po cichu, a w każdym razie lektura tego typu, przynajmniej w okresie klasycznym, była czymś bardzo rzadkim" (Chymkowski 2004: 79). Umiejętność taką miał posiadać na przykład Aleksander Wielki (Chymkowski 2004: 79). Z kolei wynalazek druku wywołał lawinę zdarzeń, których efektem było upowszechnienie umiejętności czytania na skalę masową, a dzięki lekturze – szeroki dostęp do wiedzy, szybki rozwój nauki i głębokie przemiany społeczne (Engelsing 1973, 1974).

SŁOWA KLUCZE A KORPUSY CHRONOLOGICZNE

Słowa klucze w rozumieniu lingwistycznym wyrażają najważniejsze treści występujące w konkretnym zbiorze tekstów i w pewnym zamkniętym okresie czasu. Ich generowanie z zasobów sieciowych odbywa się z ustaloną regularnością – na przykład w kolejnych dniach, tygodniach czy miesiącach. Traktowany całościowo

zbiór tekstów, z którego otrzymuje się takie dane, można określić jako k o r p u s c h r o n o l o g i c z n y. Korpus taki charakteryzuje się ścisłym i wyrażonym za pomocą metadanych uporządkowaniem sekwencyjnym tworzących go tekstów. Dzięki temu możliwe jest tworzenie rozłożonych na osi czasu histogramów częstości leksemów lub innych jednostek. Zważywszy że generowanie słów uznanych za znaczące przebiega w regularnych odstępach czasu, można przy ich opisie wykorzystać narzędzia wizualizacji i analizy szeregów czasowych[4], stosowane powszechnie w ekonomii i w naukach społecznych, a od niedawna także w lingwistyce[5].

Warto jednocześnie podkreślić, że mimo uwzględnienia zmiennej czasu korpusy chronologiczne różnią się w istotny sposób od diachronicznych. Podczas gdy istotą badań diachronicznych jest opis i wyjaśnienie z m i a n z a c h o d z ą c y c h w s y s t e m i e j ę z y k a (przede wszystkim w warstwach morfosyntaktycznej i fonetycznej), b a d a n i o m c h r o n o l o g i c z n y m p o d l e g a j ą w y ł ą c z n i e j e d n o s t k i n i e z m i e n n e p o d w z g l ę d e m b u d o w y (jednolite typograficznie i ortograficznie). Wyjątkiem od tej reguły jest zjawisko neosemantyzacji, przy którym forma leksemu może pozostać niezmienna, ewoluuje natomiast jej znaczenie – takie badania są możliwe do przeprowadzenia również w korpusach chronologicznych (problem ten zostanie omówiony w dalszej części artykułu).

Z oczywistych względów ciąg r ó ż n y c h słów kluczy wyróżnianych w następujących po sobie okresach nie tworzy szeregu czasowego – jest natomiast syntetyczną, a przez to cenną, charakterystyką przekazywanych treści. Jednak równie wartościowe wydaje się ukazanie dynamiki frekwencji konkretnych słów kluczy w długim okresie, ilustrujące stany poprzedzające moment „kluczowości" i następujące po nim. Dwa problemy poznawcze wydają się dla takiej analizy szczególnie ważne. Pierwszym jest zbadanie możliwości efektywnej selekcji potencjalnych słów tematycznych wyłącznie na podstawie metod modelowania szeregów czasowych. Badanie takie byłoby bez wątpienia cenne dla automatyzacji procesu wyszukiwawczego. Jednak kwestią poprzedzającą opracowanie rozwiązań aplikacyjnych powinno być zdefiniowanie pojęcia czasu w korpusach chronologicznych. W odróżnieniu od modeli świata fizykalnego, gdzie czas można traktować jak zmienną niezależną, w systemach społecznych jego upływ mierzy się według różnych skal.

KONCEPTUALIZACJA CZASU W BADANIACH KORPUSOWYCH

Językowa konceptualizacja czasu powstaje z połączenia apriorycznych kategorii umysłu ludzkiego (w rozumieniu kantowskim) oraz zmysłowego doświadczenia kulturowego rzeczywistości pozajęzykowej. Kategorie aprioryczne, w jakie wyposażony jest umysł człowieka, pozwalają na tworzenie spójnych, jednowymiarowych

[4] Szeregiem czasowym nazywa się w matematyce ciąg rozłożonych na osi czasu realizacji zmiennej losowej. Przez zmienną losową należy rozumieć funkcję przyporządkowującą wartości liczbowe zdarzeniom z pewnego uniwersum ontologicznego. W przypadku analizy szeregów leksykalnych (i poszukiwania słów kluczowych) oś czasu tworzą punkty wyznaczające rytm publikacji, zdarzeniami są pojawiające się w strumieniu znaków leksemy, natomiast wartościami badanej zmiennej losowej są ich częstości.

[5] Por. Glass, Wilson, Gottman 1975; Gottman 1981; Pawłowski 1998, 2001, 2005; Pawłowski, Eder 2001; Eder 2008; Pawłowski, Krajewski, Eder 2010.

i jednokierunkowych reprezentacji nieodwracalnego następstwa obserwowanych zdarzeń. Doświadczenie kulturowe jest jednak bogatsze, ponieważ oprócz linearnej, niesie ze sobą także kolistą (cykliczną) koncepcję czasu. Co ciekawe, obraz, jaki wyłania się z przeglądu kategorii czasu gramatycznego i leksykalnych wykładników relacji czasowych, wskazuje, że w systemie polszczyzny owo wielowymiarowe doświadczenie modelowane jest niemal wyłącznie jako proces linearny (por. Grzegorczykowa 2016; Sawicka 2006; Rokoszowa 1998). Także badania historycznojęzykowe są zanurzone w „podłużnym" kontinuum czasu. Jednym z wyjątków od tej reguły jest aspekt iteracyjny czasownika, który środkami morfologicznymi (a nie semantycznymi) wyraża cykliczność.

W badaniach korpusowych czas jest tradycyjnie kojarzony z podziałem na lingwistykę synchroniczną i diachroniczną, ukształtowanym w pierwszej połowie dwudziestego wieku. Jego pochodną jest rozróżnienie korpusów synchronicznych i diachronicznych. Jak wyżej wspomniano, te ostatnie zawierają teksty z różnych, nawet bardzo odległych okresów historycznych, pozwalające dostrzec ewolucję języka. Natomiast korpusy chronologiczne złożone są z próbek tekstów uporządkowanych na osi czasu według dat wydania lub powstania, a przy tym jednolitych pod względem ortograficznym i typograficznym. W obu wypadkach czas, przynajmniej z pozoru, ma linearną strukturę wektora biegnącego nieodwracalnie od przeszłości ku przyszłości.

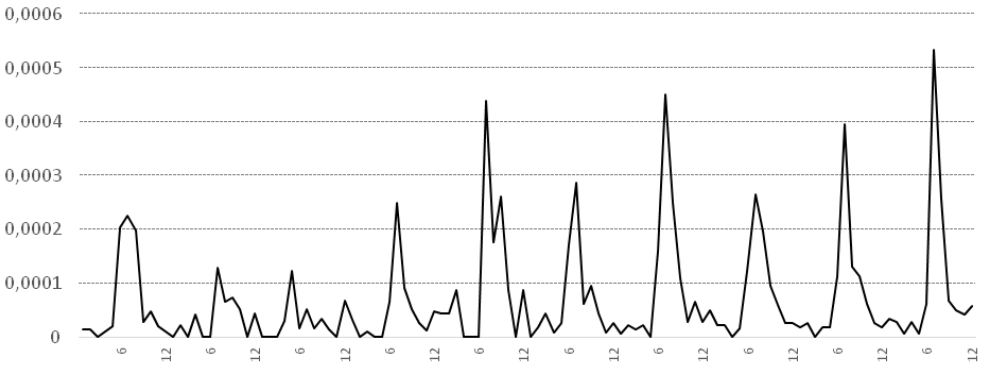
Jednak ta powszechnie stosowana koncepcja czasu jest dość upraszczająca i nie wystarcza do wyjaśnienia wszystkich histogramów generowanych z danych zawartych w korpusach chronologicznych. Bardziej wnikliwe podejście pozwala wyróżnić trzy porządki pojęciowe, dzięki którym możliwe jest uchwycenie bogactwa procesów i zjawisk, jakie kryją się za pozornie prostymi szeregami frekwencji leksemów w kolejnych próbkach tekstowych korpusu chronologicznego. Porządki te proponuję określić jako strukturalny, antropologiczny i metodologiczny.

Porządek strukturalny obejmuje wspomniane już bazowe kategorie czasu linearnego (czyli wektora biegnącego nieodwracalnie w kierunku umownie zwanym przyszłością) oraz kolistego (czyli powtarzalnych w nieskończoność cykli tych samych zjawisk, zdarzeń, procesów itd.). Porządek antropologiczny wyraża naturalną bądź kulturową genezę zjawisk, jakie kryją się za histogramami frekwencji leksemów. W grupie tej można wyróżnić kategorie czasu astronomicznego, kulturowego, politycznego i cywilizacyjnego. Trzy pierwsze mają w okresach krótkich (maksymalnie kilkudekadowych) strukturę kolistą, natomiast perspektywa dłuższa (na przykład stulecia) może ukazać także linearność, charakterystyczną dla czasu cywilizacyjnego. Wreszcie porządek trzeci – metodologiczny – obejmuje formalne (matematyczne) reprezentacje czasu w porządkach antropologicznym i strukturalnym.

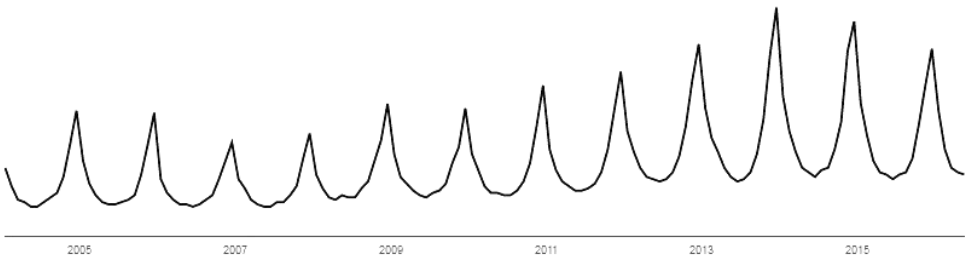
Kategorie strukturalne (linearność, kolistość) są intuicyjnie jasne i w tym kontekście służą jedynie do opisu konceptualizacji czasu postrzeganego jako element przestrzeni społecznej. Najwartościowszy składnik poznawczy w analizie szeregów leksykalnych stanowią bez wątpienia kategorie należące do porządku antropologicznego[6], czyli tej odmiany czasu, w której zanurzone są długotrwałe procesy społeczne, kulturowe i gospodarcze.

[6] Ogólne pojęcie czasu antropologicznego i rozbudowany system jego podkategorii są omówione m.in. w pracach Małgorzaty Krzysztofik (2010, 2013). Wiele cennych spostrzeżeń na temat relatywizmu czasu zawiera także artykuł Andrzeja Sicińskiego (1975).

Czas astronomiczny determinowany jest regularnym ruchem Ziemi i ciał niebieskich. Reguluje on aktywność człowieka związaną z cyklami klimatycznymi, a w szczególności prace rolnicze oraz inne zjawiska zależne od pogody, do których zaliczyć można m.in. sezonowe epidemie (np. grypę), czynności administracyjne (np. odśnieżanie, szczepienia profilaktyczne), zachowania lub aktywności publiczne (sposób ubierania się, uprawianie sportu), a nawet specyficzne wypadki występujące tylko w pewnych porach roku (np. utonięcia). Korpusy tekstów rejestrują w określonych momentach regularne zwiększanie się częstości występowania słownictwa związanego z tymi właśnie obszarami tematycznymi. Jako przykłady takiej charakterystyki chronologicznej przedstawiono histogramy leksemów *żniwa* i *scarf* (szalik), które w oczywisty sposób zależą od czasu astronomicznego (rys. 1, 2).



Rys. 1. Histogram rzeczownika *żniwa* w korpusie ChronoPress, osadzony w czasie astronomicznym (znormalizowane frekwencje za lata 1945–1954)[7]

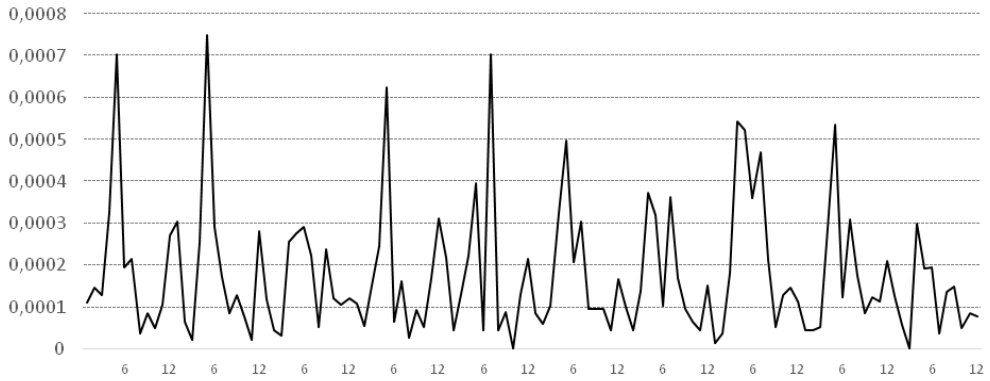


Rys. 2. Histogram znormalizowanych częstości rzeczownika *scarf* (szalik) w korpusie Google Books[8]

[7] Źródło: http://chronopress.clarin-pl.eu/ [dostęp 25.03.2016]. Na osi odciętych przedstawiono kolejne miesiące lat 1945–1954, a na osi rzędnych częstość względną leksemu.

[8] Grafika ze strony narzędzia Google Trends: https://www.google.com/trends/ [dostęp 30.03.2016]. Nie podano dokładnej metody wyliczania znormalizowanych frekwencji leksemów na osi rzędnych. Wszystkie kolejne histogramy generowane z zasobów Google Books pochodzą z cytowanej strony WWW i były pozyskane w tym samym momencie.

Czas kulturowy reguluje zjawiska, które wynikają z powtarzalności świąt (świeckich i religijnych), rocznic i innych rytuałów kulturowych o charakterze prywatnym lub publicznym (można zaliczyć do tej grupy także cykliczne imprezy sportowe). Także w tym wypadku mamy do czynienia z systematycznym pojawianiem się wysokiej częstości słownictwa powiązanego z zdarzeniami kulturowymi. Rysunek 3 przedstawia znormalizowaną częstość leksemu *święto*, który osiąga swoje regularne ekstrema w kwietniu i w maju każdego roku, a dzieje się tak dlatego, że oficjalna prasa realizowała program ateizacji społeczeństwa, polegający m.in. na pomijaniu świąt kościelnych, zamiast których eksponowano „święta świeckie" – wśród których najważniejszy był pierwszy maja (rys. 3). Czas polityczny jest determinowany regulacjami prawnymi, które określają terminy wyborów, posiedzeń ciał kolektywnych, ogłaszania aktów prawnych, celebracji rocznic państwowych, kampanii informacyjnych lub propagandowych etc. Podobnie jak w przypadku czasu kulturowego aktywność polityczna przejawia się powtarzalnym wzrostem frekwencji stosownego słownictwa (por. Borowiec 2013).



Rys. 3. Histogram rzeczownika *święto* w korpusie ChronoPress, osadzony w czasie kulturowym (frekwencje względne za lata 1945–1954)[9]

Wymienione tutaj cykle astronomiczne, kulturowe i polityczne są trwałe i można uznać je za fundament stabilności społecznej, a pośrednio źródło poczucia bezpieczeństwa. Różnica między koncepcjami czasu kulturowego i politycznego polega głównie na tym, że zdarzenia polityczne są regulowane prawem, natomiast cykle kulturowe mają charakter mniej formalny, ale za to utrwalony obyczajem. Warto przy tej okazji zauważyć, że tzw. cykle gospodarcze, chętnie przywoływane w teorii ekonomii, wydają się trudne do wychwycenia jedynie na podstawie frekwencji określonych leksemów. Aby je dostrzec, można na przykład rozważyć pary terminów antonimicznych *hossa*, *bessa*, *rozwój*, *kryzys* i ich synonimów. Jednak problemem podstawowym jest sama cykliczność zjawisk gospodarczych, która wydaje się być bardziej postulatem lub samospełniającą się przepowiednią niż obiektywną rzeczywistością[10].

[9] Źródło: http://chronopress.clarin-pl.eu/ [dostęp 25.03.2016]. Na osi odciętych przedstawiono kolejne miesiące lat 1945–1954, a na osi rzędnych częstość względną leksemu.

[10] „Teoria", dla której punktem wyjścia jest biblijna przypowieść o siedmiu krowach albo spekulatywne wątki filozofii pitagorejskiej, pozostaje przynajmniej w części domeną wiary, a nie nauki.

Czwarty rodzaj czasu, należący do porządku antropologicznego, określony tutaj jako czas cywilizacyjny, jest zjawiskiem tak ważnym, że należy poświęcić mu więcej miejsca. Jego koncepcja wyrasta z teorii długiego trwania (*longue durée*), kojarzonej z francuską szkołą Annales i pracami jej przedstawicieli (Braudel 1958; por. także Sowa 2008: 130–132). Zgodnie z tą teorią procesy historyczne mają podłoże ekonomiczno-społeczne i rozwijają się przez wieki, a zdarzenia jednorazowe o charakterze anomalii (na przykład wojny, epidemie, katastrofy naturalne, zmiany dynastyczne) mogą takie trwałe trendy kulturowe jedynie chwilowo zaburzać, ale nie powstrzymywać. W tym duchu (chociaż wiele lat wcześniej) wypowiadał się antropolog kultury Norbert Elias. Dostrzegał on trwającą przez wieki powolną ewolucję kulturowych manifestacji popędów, której początku należy szukać w głębokim średniowieczu: „Jakkolwiek więc przebiegają, zwłaszcza obserwowane z bliskiej perspektywy, owe krzyżujące się ruchy, owe przypływy i odpływy fal niosących już to zacieśnienie, już to rozluźnienie więzów, kierunek główny ewolucji – tak jak rysuje się ona dotąd – jest ten sam, bez względu na rodzaj popędów i ich manifestacji. Krzywa ewolucji popędu płciowego, ogólnie biorąc, przebiega w procesie cywilizacji równolegle do krzywych innych manifestacji popędów, jeśli nawet w szczegółach zachodzić mogą między nimi różnice socjogenetyczne" (Elias 1980: 268). Ta wyobrażona „krzywa ewolucji popędów" ma niewątpliwie swoją werbalną reprezentację, którą, dysponując odpowiednimi korpusami i nie bez trudności typowych dla analiz semantycznych, dałoby się wyrazić w postaci szeregu leksykalnego wygenerowanego ze strumienia danych tekstowych. W podobnym duchu Marek Łaziński wyraża się o grzecznościowej zasadzie „ograniczania bezpośredniości", którą przedstawia jako długotrwały, wielopokoleniowy i jednokierunkowy proces stopniowego oddalania się nazw poprawnych politycznie od cech, które w przeszłości były podstawą stygmatyzacji (Łaziński 2014: 131). Owo ograniczanie bezpośredniości byłoby więc jednym z trendów cywilizacyjnych i kulturowych, które można badać w perspektywie chronologicznej.

Przenosząc te rozważania na plan metodologiczny analizy korpusowej, można przyjąć jako bardzo prawdopodobną hipotezę, że znormalizowana frekwencja względna[11] każdego leksemu w dostatecznie długim okresie czasu układa się w trend rosnący lub malejący, chociaż w krótkich odcinkach może wydawać się stała lub podlegać wyłącznie oscylacjom periodycznym i fluktuacjom losowym (Pawłowski 2006). Trendy takie wyrażają różne procesy zachodzące właśnie w zdefiniowanym wyżej czasie cywilizacyjnym, dostrzeganym dzięki korpusom chronologicznym. Teza powyższa została potwierdzona licznymi badaniami materiału pochodzącego z korpusu ChronoPress, a także poglądowymi analizami zasobów online, jakie można wygenerować narzędziami Google Ngram Viewer i Google Analytics (z uwagi na ograniczoną objętość artykułu przedstawiono tylko niektóre z tych analiz).

Odnosząc się wreszcie do porządku metodologicznego, należy powiedzieć, że obejmuje on formalne reprezentacje szeregów leksykalnych. Przyjmuje się, że szeregi czasowe (także te, które generuje się z korpusów chronologicznych)

[11] Jest to przeliczeniowa częstość leksemu, odpowiadająca stałej objętości korpusu (na przykład sto tysięcy lub milion tokenów). Warto pamiętać, że normalizacja polegająca na proporcjonalnym przeliczeniu częstości jest jedynie uproszczeniem, ponieważ między frekwencją leksemu a wielkością próby zależność jest nieliniowa.

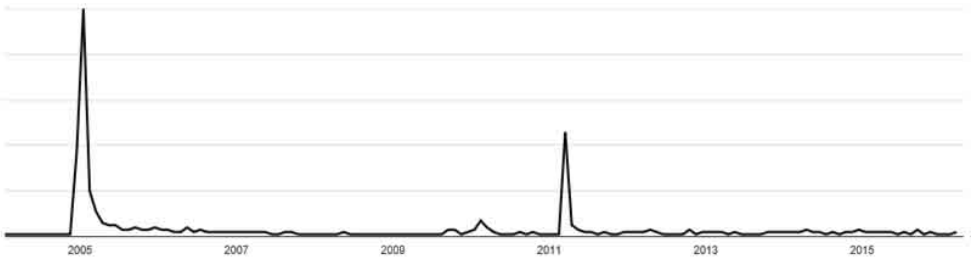
zawierają trzy składowe: długotrwały trend monotoniczny, oscylacje periodyczne i wahania losowe. Każda składowa może być reprezentowana odpowiednią funkcją matematyczną, można także analizować współzależność wielu szeregów jednocześnie, stosując funkcję regresji. Czas cywilizacyjny byłby modelowany jako trend monotoniczny (struktura linearna), natomiast czasy astronomiczny, kulturowy i polityczny (struktura kolista) jako oscylacje periodyczne. Taka postać modelu jest zgodna z teorią szeregów czasowych w wersji powszechnie wykorzystywanej w naukach stosowanych (przede wszystkim w ekonometrii). Podczas analizy szeregów leksykalnych okazało się jednak, że model ten jest niewystarczający, ponieważ pomija pewne wartościowe pod względem poznawczym przypadki.

SZEREGI LEKSYKALNE A ROZPOZNAWANIE SŁÓW KLUCZY

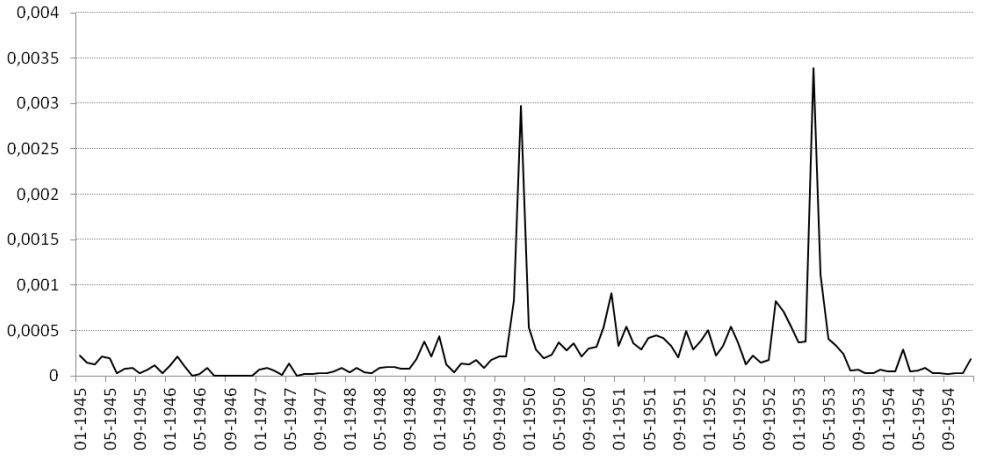
Wśród zjawisk społecznych opisywanych przedstawionymi wyżej kategoriami czasu i modelami szeregów leksykalnych dominują procesy w jakimś stopniu systematyczne, przewidywalne, pozwalające na ekstrapolację i krótkookresową predykcję. Ale od zarania dziejów człowiekowi towarzyszyły także zdarzenia nagłe, intensywne i nieprzewidywane – jednym słowem anomalie. Były one przez długie wieki konceptualizowane w kategoriach nieuchronnego przeznaczenia, metafizycznej kary lub boskiej interwencji (utrwalone kulturowo wzorce takich rozumowań znaleźć można m.in. w mitologii greckiej i w tekstach starotestamentowych). Jako zagrożenie dla ustalonego porządku rzeczy anomalie podlegały częściowemu wyparciu z przestrzeni komunikacyjnej. Dopiero rozwój przemysłu informacyjnego, opartego na zaspokajaniu masowych potrzeb, uczynił z katastrof, skandali i innych nieprzewidywalnych zdarzeń preferowany przez publiczność czytelniczą przedmiot mediatyzacji. Warto przy tej okazji podkreślić, że współczesne opisy medialne tylko pozornie zdają relację z „rzeczywistych wydarzeń", lecz w istocie konstruują niemal autonomiczną rzeczywistość komunikacyjną.

Chcąc wprowadzić do obszaru badań korpusowych tekstowe reprezentacje zjawisk katastroficznych i innych anomalii, należy przyjąć, że nie ma dla nich specjalnej odmiany czasu antropologicznego. Pojawiają się one bowiem nie jako procesy, lecz jako jednorazowe i krótkotrwałe zaburzenia cykli periodycznych lub trendów. Można natomiast rozszerzyć listę podstawowych modeli szeregów leksykalnych o w z o r z e c a n o m a l i i, wyrażający gwałtowny wzrost lub spadek wartości częstości leksemu. Należy przy tym zauważyć, że nie istnieje specjalna klasa wyrazów, które realizowałyby ten schemat. Na przykład leksemy *pożar*, *awaria*, *eksplozja*, *zamach* czy *śmierć*, podobnie jak nazwy własne powiązane z sytuacjami katastroficznymi, mogą być rozłożone w całym korpusie chronologicznym równomiernie i z przeciętną częstością. Dopiero na skutek zdarzeń zewnętrznych o intensywnym oddźwięku społecznym i medialnym ich częstość w tekstach będzie w jakimś momencie gwałtownie wzrastać. Przykładem takiego zachowania jest histogram rzeczownika *tsunami* (rys. 4) oraz nazwy własnej *Stalin* (rys. 5). Na uwagę zasługuje obecność na rysunku 5 dwóch ekstremów o charakterze „katastroficznym": drugie związane jest oczywiście ze śmiercią przywódcy ZSRR w marcu 1953 roku, natomiast pierwsze, zaskakująco silne, jest śladem pojawienia się na przełomie

1949 i 1950 roku wielkiej liczby czołobitnych artykułów z okazji siedemdziesiątej pierwszej rocznicy jego urodzin.



Rys. 4. Histogram znormalizowanych częstości rzeczownika *tsunami* w korpusie Google Books[12]



Rys. 5. Histogram znormalizowanych częstości rzeczownika *Stalin* w korpusie ChronoPress[13]

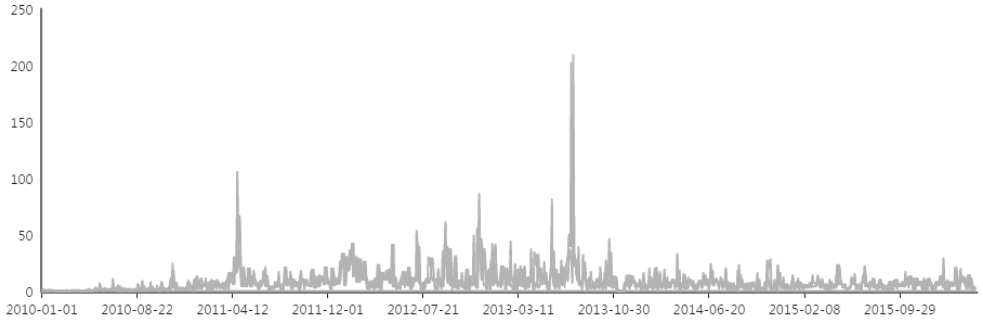
Właśnie wśród wyrazów charakteryzujących się bardzo wysoką i niedającą się przewidzieć dynamiką zmian częstości należy szukać potencjalnych słów kluczy. Na użytek analizy szeregów leksykalnych można jednostki takie określać mianem w y r a z ó w - k o m e t. Użyta tutaj metafora komety pojawia się nieprzypadkowo. Zgodnie z tradycyjnym oglądem wszechświata niebo było obszarem porządku, gdzie pewne punkty pozostawały niezmienne (na przykład gwiazdozbiory) lub poruszały się cyklicznie po ustalonych torach. Kosmos, grecka nazwa wszechświata, pochodzi od rzeczownika greckiego *κόσμος*, oznaczającego właśnie *porządek*. Kometa była więc, zgodnie z tym tradycyjnym podejściem, zjawiskiem nieprzewidywanym, niewytłumaczalnym, a więc anomalią. Jeżeli przyjąć, że dyskurs publiczny jest w miarę stabilnym strumieniem danych, chociaż pulsującym swoim wewnętrznym rytmem, gwałtowny i nieprzewidywalny skok częstości jakiegoś leksemu byłby właśnie meta-

[12] Źródło: Google Trends: https://www.google.com/trends/ (zrzut ekranowy z 25.03.2016). Nie podano dokładnej metody wyliczania znormalizowanych frekwencji leksemów na osi rzędnych.

[13] Źródło: http://chronopress.clarin-pl.eu/ [dostęp 25.03.2016]. Na osi odciętych przedstawiono kolejne miesiące lat 1945–1954, a na osi rzędnych częstość względną leksemu.

forycznym odpowiednikiem pojawienia się komety – także zaburzającej stan stabilny i przewidywalny.

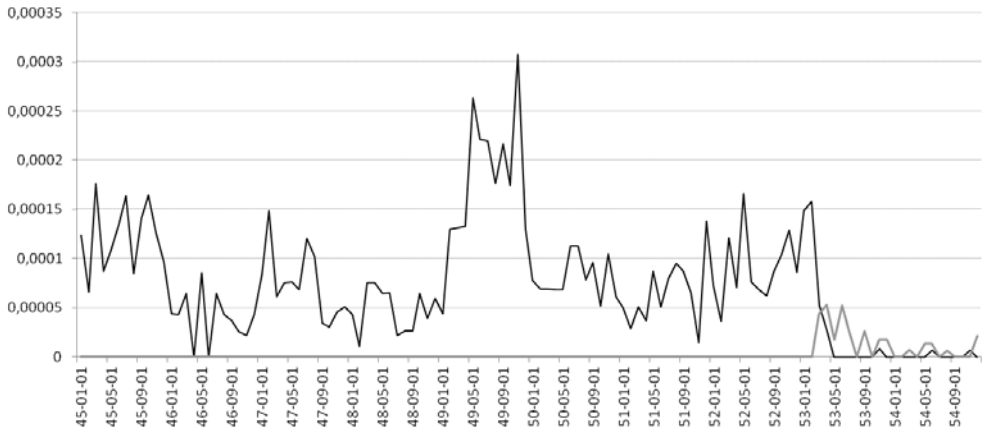
Tak jak wyżej wspomniano, w zasadzie każdy leksem może znaleźć się w klasie potencjalnych wyrazów-komet, a w konsekwencji słów kluczy. Badania empiryczne wykazały jednak, że pole semantyczne katastrofy – o ile takowa wystąpi w rzeczywistości pozajęzykowej – jest reprezentowane ze szczególną intensywnością. Co ciekawe funkcję tę mogą pełnić również leksemy o nacechowaniu neutralnym lub pozytywnym, reprezentujące dowolny mediatyzowany temat. Rysunek 6 przedstawia na przykład znormalizowane frekwencje nazwy własnej *Kate*, odnoszącego się do członkini brytyjskiej rodziny królewskiej, z której media całego świata uczyniły bohaterkę wyobraźni kolektywnej w dwóch globalnych kampaniach promocyjnych (rys. 6). Osobną grupę leksemów, które mogą spełnić formalne kryteria wyrazów-komet, są niektóre jednostki słownikowe podlegające oscylacjom periodycznym. Periodyczność ma bowiem tę cechę, że widoczna jest dopiero w dłuższej perspektywie: na przykład przy sezonowości rocznej (najczęściej wstępującej) jej zaobserwowanie wymaga uwzględnienia danych przynajmniej z kilku okresów dwunastomiesięcznych. Skrócenie badanego okresu do roku może natomiast sprawić, że, na przykład, gwałtowny wzrost częstości leksemu *grypa* w okresie jesiennym lub *alergia* na wiosnę będzie odebrany nie jako przewidywalny ruch krzywej periodycznej, lecz rodzaj anomalii.



Rys. 6. Histogram znormalizowanych częstości nazwy własnej *Kate* w zasobach sieci Internet[14]

Dodatkowo należy zwrócić uwagę na to, że kategoria słowa klucza wymaga intensywnego wzrostu częstości danego leksemu, podczas gdy wzorzec anomalii mieści w sobie także sytuację odwrotną, czyli jej nagły spadek. Do grupy tej należy na przykład nazwa własna *Katowice*, która po zmianie nazwy tego miasta na *Stalinogród* w roku 1953 zniknęła na kilka lat z piśmiennictwa oficjalnego (rys. 7). Podobnej charakterystyki należałoby oczekiwać w przypadku nazwy własnej *Tito*, czyli przydomka komunistycznego przywódcy powojennej Jugosławii, który w 1949 roku wypadł z łask propagandy sowieckiej i zniknął z dyskursu publicznego państwa bloku wschodniego jako „przyjaciel". Tutaj jednak sam szereg czasowy nie jest wystarczającym źródłem informacji, ponieważ po krótkiej nieobecności *Tito* powrócił jako figura wroga (dane nie zostały tutaj przedstawione ze względu na zbyt niskie frekwencje i małą wyrazistość histogramu).

[14] Źródło: wyszukiwarka Frazeo.pl (zrzut ekranowy wykonano 25.03.2016).



Rys. 7. Histogram znormalizowanych częstości nazw własnych *Katowice* i *Stalinogród* w korpusie ChronoPress[15]

WYKRYWANIE NEOSEMANTYZACJI ZA POMOCĄ SZEREGÓW LEKSYKALNYCH

Większość istniejących obecnie korpusów tekstów powstawała w ramach projektów leksykograficznych – jako reprezentacja „prawdziwego języka" miały one służyć do ekscerpcji haseł i przykładów ich użycia. Jednym z zadań o takim charakterze jest komputerowo wspomagane rozpoznawanie nowych jednostek w systemie języka, czyli neologizmów i neosemantyzmów.

Stosowane w praktyce metody automatycznego wykrywania neologizmów można podzielić na lingwistyczne, teoriomnogościowe i statystyczne (por. Janicijevic, Walker 1997; Janssen 2005, 2009; Wierzchoń 2005; Paryzek 2008; Kolkovska et al. 2012).

Podejście lingwistyczne opiera się na założeniu, że nadawca komunikatu opatruje każdy nowy wyraz znacznikami metatekstowymi, antycypując problemy odbiorcy z jego zrozumieniem. Na przykład dla języka angielskiego Piotr Paryzek uznał za wyznaczniki potencjalnie nowych form wyrażenia *called*, *defined as*, *known as*, *termed* oraz cudzysłów (cf. Paryzek 2008). Zaletą tego rozwiązania jest odwołanie się do świadomości językowej konkretnych użytkowników. Jednak to właśnie jest równocześnie potencjalnym źródłem rozbieżnych ocen, ponieważ ta sama forma może być odbierana przez jednych jako nieznana, a przez drugich jako już istniejąca.

Podejście teoriomnogościowe opiera się na założeniu, że korpusy są zbiorami elementów leksykalnych w sensie matematycznym i można wykonywać na nich operacje dodawania, mnożenia i odejmowania. Zadanie wykrycia neologizmu polega na znalezieniu różnicy zbiorów reprezentujących korpusy referencyjny i badany (cf. Kolkovska et al. 2012). W tym celu generuje się i porównuje listy jednostek, wyszukując elementy nowe, niewystępujące w korpusach referencyjnych

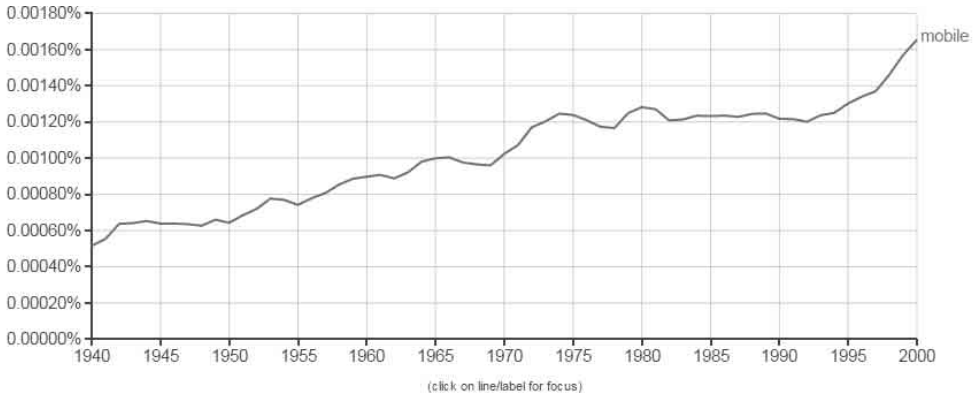
[15] Źródło: http://chronopress.clarin-pl.eu/ [dostęp 25.03.2016].

(można w uproszczeniu powiedzieć, że korpus referencyjny jest rodzajem stoplisty dla korpusu badanego). Podejście takie ma zaletę polegającą na wykorzystaniu nieograniczonych możliwości obliczeniowych komputera. Jego wadą jest natomiast „spłaszczenie" języka do warstwy graficznej. Powołując się m.in. na prace Haralda Baayena i Antoinette Renouf, Maarten Janssen wspomina jeszcze o podejściu statystycznym, które opiera się na prawidłowości zaobserwowanej w badaniach empirycznych, polegającej na tym, że potencjalne neologizmy pojawiają się początkowo jako jednostki z frekwencją równą jedności (*hapax legmena*) (Janssen 2009: 71).

Zmiany leksykalne nie ograniczają się jednak do tworzenia zupełnie nowych jednostek. Oprócz neologizacji istnieje przecież częściej spotykane zjawisko neosemantyzacji, czyli modyfikacji bądź rozszerzenia znaczenia istniejącego już leksemu. Automatyczne wykrywanie neosemantyzmów jest trudne, ponieważ odwołuje się do warstwy znaczeniowej języka, a nie do rozpoznawalnych maszynowo form. Dlatego intuicja językowa, niemal bezbłędna w wykrywaniu nowych znaczeń leksemów, jest w tym wypadku zawsze ostateczną instancją oceniającą. Praca człowieka jest jednak powolna i kosztowna, co sprawia, że czas wymagany do przetworzenia nawet średniego korpusu tekstów jest niewspółmierny do osiągniętego rezultatu. Istnieją jednak metody badawcze ułatwiające automatyczne rozpoznawanie potencjalnych neosemantyzmów w dużych korpusach tekstów: należy do nich analiza dystrybucji leksemu.

Można z dużą dozą pewności przyjąć, że każda zmiana lub rozszerzenie znaczenia leksemu ma wpływ na jego dystrybucję (łączliwość w tekstach). Widać to wyraźnie w „otoczeniu" takich wyrazów, jak *aplikacja*, *fura*, *mysz*, *program*, *promocja* czy *zacny* w różnych okresach historycznych. Jednak warunkiem rozpoznania neosemantyzacji tą metodą jest możliwość wygenerowania z korpusu poprawnego zbioru kolokatów. W języku polskim jest to utrudnione z uwagi na fakt, że do wskazania jednostek rzeczywiście powiązanych wymagane jest efektywne oznaczenie struktury składniowej zdań (obecnie rozwiązanie tego problemu jest jeszcze w fazie badawczej – por. Patejuk, Przepiórkowski 2015). W praktyce stosuje się natomiast mechaniczny pomiar łączliwości na podstawie współwystępowania w linii tekstu – łatwy z programistycznego punktu widzenia, ale obciążony błędem. Na przykład na podstawie zdania *Obiad będzie wieczorem zimny* mechaniczny kolokator wykaże, wbrew intuicji językowej, współwystępowanie leksemów *obiad* i *zimny*. Poprawna analiza tego zdania wymagałaby odróżnienia frazy głównej od okolicznikowej, co nie jest możliwe bez pełnego oznaczenia składniowego.

Można jednak zaproponować i poddać testom rozwiązanie inne, niewymagające analizy kolokacji, odwołujące się natomiast do analizy szeregów leksykalnych w formie zbliżonej do automatycznego wyznaczania słów kluczy. Dotychczasowe badania wykazały, że w długim okresie, w kolejnych segmentach czasowych, istnieje znacząca korelacja między dynamiką zmian częstości leksemu a zmianą lub rozszerzeniem jego zakresu znaczeniowego. Zgodnie z tą prawidłowością każda trwała zmiana trendu, jaki wyznaczają częstości leksemu, może wskazywać na potencjalną neosemantyzację. Na przykład przedstawiony na rysunku 8 histogram częstości wyrazu angielskiego *mobile*, wyznaczony na podstawie danych Google Books, notuje kilka zmian kierunku trendu, przy czym najważniejsza wydaje się ta z początku lat dziewięćdziesiątych, kiedy najprawdopodobniej znaczenie przymiotnikowe (*ruchomy*) zaczęło ustępować rzeczownikowemu (*telefon komórkowy*).



Rys. 8. Szereg oparty na częstościach leksemu *mobile* (początkowo przymiotnik *ruchomy*, od końca lat 80. *telefon komórkowy*)[16]

Jednak badania prowadzone na materiale korpusów ChronoPress, Google Books oraz wyszukiwarki Frazeo, indeksującej otwarte zasoby sieci WWW (http://frazeo.pl/), wykazały, że metoda wykrywania neosemantyzmów na podstawie zmian histogramu jest efektywna dopiero na dużych korpusach chronologicznych, obejmujących przynajmniej ćwierćwiecze. Niestety zasoby takie dla większości języków nie zostały jeszcze wytworzone. Na przykład ChronoPress obejmuje dopiero jedną dekadę i zawiera próby o objętości około stu tysięcy wyrazów na miesiąc, co uniemożliwia odkrywanie trendów długoterminowych i badanie leksemów o niewielkich częstościach. Łatwiejsze okazało się natomiast opracowanie metody automatycznego rozpoznawania histogramów o wymaganym kształcie (tutaj – zawierających trend i jego zmianę). Testy wykazały, że jeżeli wartość współczynnika korelacji wzajemnej (*cross-correlation*) szeregów wzorcowego i badanego ma wartość powyżej 0,6, można mówić o ich znaczącym podobieństwie.

WNIOSKI

Konieczne wydaje się wprowadzenie do repertuaru pojęć lingwistyki korpusowej kategorii korpusu chronologicznego, definiowanego w opozycji do synchronicznego i diachronicznego (historycznego). Dzięki nowym aplikacjom i zasobom badanie szeregów leksykalnych staje się coraz łatwiejsze i zyskuje na popularności. Dowodzą tego rozwijające się serwisy internetowe (przede wszystkim Google Ngram Viewer, Frazeo, niektóre funkcjonalności korpusów narodowych), ale także zasoby dedykowane (na przykład korpus ChronoPress).

W badaniach szeregów leksykalnych warto wykorzystywać antropologiczne kategorie czasu astronomicznego, kulturowego, politycznego i cywilizacyjnego, pozwalające na lepsze zrozumienie zjawisk kryjących się za histogramami leksemów.

[16] Grafika ze strony narzędzia Google Ngram Viewer: https://books.google.com/ngrams/ [dostęp 9.03.2016].

Podstawowe kategorie metodologiczne, wykorzystywane w badaniu korpusów chronologicznych, pochodzą z teorii szeregów czasowych. Do repertuaru podstawowych wzorców (trend, oscylacje, wahania losowe) należy jednak dołożyć wzorzec anomalii, pozwalający na rozpoznawanie zjawisk gwałtownych i nieprzewidywalnych. Istotnym wnioskiem z przeprowadzonych badań jest również możliwość wykorzystania szeregów leksykalnych do generowania słów kluczy, a także do rozpoznawania neosemantyzmów. Potencjalne słowa klucze należałyby do kategorii tzw. wyrazów-komet, natomiast neosemantyzacja byłaby rozpoznawana dzięki zmianie trendu.

Jeżeli uznać prasę za najlepsze źródło danych do badań chronologicznych, podstawowym ograniczeniem w badaniach, przynajmniej w odniesieniu do polszczyzny, jest niewielka skala dostępności opracowanego cyfrowo materiału tekstowego[17]. W przeciwieństwie do tekstów nowych, powstających od razu w postaci cyfrowej, wytworzenie zasobów historycznych prasy w postaci czytelnej dla komputera zabiera całe lata i jest bardzo kosztowne. Za podstawowy kierunek prac nad korpusami chronologicznymi należy więc uznać przede wszystkim ich rozbudowę w oparciu o periodyki publikowane od roku 1800 do współczesności. Oprócz tego konieczne jest dopracowanie stosowanego obecnie formatu metadanych oraz wykorzystanie stworzonych w ramach Clarin-PL i innych projektów narzędzi automatycznego przetwarzania języka.

BIBLIOGRAFIA

Augustyn (św.) (1987), *Wyznania*. Przełożył, opatrzył posłowiem i kalendarium Zygmunt Kubiak. Warszawa: Pax.

Babik Wiesław (2010), *Słowa kluczowe*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Bojar Bożenna (red.) (2002), *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*. Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

Borin Lars, Forsberg Markus, Roxendal Johan (2012), Korp – the corpus infrastructure of Språkbanken. Proceedings of LREC 2012. Istanbul: ELRA, 474–478.

Borowiec Piotr (2013), *Czas polityczny po rewolucji*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Braudel Fernand (1958), Histoire et Sciences sociales: La longue durée. *Annales. Économies, Sociétés, Civilisations* no 4, 725–753.

Brunet Etienne (1981), *Le vocabulaire français. De 1789 à nos jours*. Paris–Genève: Slatkine, Champion.

Castells Manuel (2007), *Społeczeństwo sieci*. Przekład i red. naukowa Mirosława Marody [i in.]. Warszawa: Wydawnictwo Naukowe PWN.

Chmielewska-Gorczyca Ewa (1995), Funkcje tezaurusa w systemie informacyjno-wyszukiwawczym. *Zagadnienia Informacji Naukowej* 1–2, 3–17.

Chymkowski Roman (2004), *Głośna i cicha lektura w starożytnej Grecji*. W: Katarzyna Lange, Władysław Sawrycki, Paweł Tański (red.), *Problematyka tekstu głosowo interpretowanego* 2. Toruń: Wyd. Adam Marszałek, 75–83.

[17] Dane chronologiczne m.in. z tekstów prasowych, obejmujące okres od roku 1800, zawiera na przykład korpus skandynawski Korp Språkbanken (Borin, Forsberg, Roxendal 2012; por. też Viklund, Borin 2016: 86).

Cortada James W. (2002), *Making the information society: experience, consequences, and possibilities*. Upper Saddle River, NJ: Prentice Hall.

Cortada James W. (2012), *The Digital Flood*. New York: Oxford University Press.

Eder Maciej (2008), How rhythmical is hexameter: a statistical approach to Ancient epic poetry. *Digital Humanities* 2008, Book of Abstracts. University of Oulu, 112–114.

Elias Norbert (1980), *Przemiany obyczajów w cywilizacji Zachodu*. Przełożył Tadeusz Zabłudowski. Warszawa: Państwowy Instytut Wydawniczy.

Engelsing Rolf (1973), *Analphabetentum und Lektüre. Zur Sozialgeschichte des Lesens in Deutschland zwischen feudaler und industrieller Gesellschaft*. Stuttgart: Metzler.

Engelsing Rolf (1974), *Der Bürger als Leser: Lesergeschichte in Deutschland von 1500–1800*. Stuttgart: Metzler.

Genette Gérard (1987), *Seuils*. Paris: Éditions du Seuil.

Glass Gene V., Wilson Victor L., Gottman John M. (1975), *Design and Analysis of Time-Series Experiments*. Colorado: Colorado Associated University Press.

Gottman John M. (1981), *Time-series analysis: a comprehensive introduction for social scientists*. Cambridge, London etc.: Cambridge University Press.

Góralska Małgorzata (2012), *Piśmienność i rewolucja cyfrowa*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

Grzegorczykowa Renata (2006), *Polskie przymiotniki temporalne na tle ogólnej językowej konceptualizacji czasu*. W: Anna Dąbrowska, A. Nowakowska (red.), *Czas, język, kultura. Język a kultura* 19, 33–44.

Havelock Eric A. (2006), *Muza uczy się pisać: rozważania o oralności i piśmienności w kulturze Zachodu*. Przekł. i wstęp Paweł Majewski. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Janicijevic Tatjana, Walker Derek (1997), NeoloSearch: Automatic Detection of Neologisms in French Internet Documents. *ACH-ALLC'97*, Kingston, Canada.

Janssen Maarten (2005), NeoTrack: semiautomatic neologism detection. Associação Portuguesa de Linguística, *APL* XXI, Porto. http://maarten.janssenweb.net/Papers.

Janssen Maarten (2009), Detección de Neologismos: una perspectiva computacional. *Debate Terminológico* 5, 68–75.

Kolkovska Sia, Blagoeva Diana, Atanasova Atanaska (2012), The application of corpus-based approach in the Bulgarian new-word lexicography. In: Ruth Vatvedt Fjeld, Julie M. Torjusen (eds.), Proceedings of the 15th EURALEX International Congress. 7–11 August 2012. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 991–996. http://www.euralex.org/proceedings-toc/euralex_2012.

Krzysztofik Małgorzata (2010), *Studium z dziejów krakowskich kalendarzy astrologicznych XVII wieku. Almanachy Stanisława Słowakowica jako podstawa uogólnień*. Kraków: Księgarnia Akademicka.

Krzysztofik Małgorzata (2013), *Kategoria czasu antropologicznego w polskim kalendarzu XVII-wiecznym*. W: Iwona M. Dacka-Górzyńska, Joanna Partyka (red.), *Kalendarze staropolskie*. Warszawa: Wydawnictwo DiG, 61–87.

Lem Stanisław (1999), *Bomba megabitowa*. Kraków: Wydawnictwo Literackie.

Loewe Iwona (2007), *Gatunki paratekstowe w komunikacji medialnej*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Łaziński Marek (2104), Jeszcze raz o słowie *murzyn* i o stereotypach. Po lekturze artykułu Margaret Ohii *Mechanizm dyskryminacji rasowej w systemie języka polskiego*. *Przegląd Humanistyczny* 5, 127–141.

Malak Piotr (2012), *Indeksowanie treści: porównanie skuteczności metod tradycyjnych i automatycznych*. Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

Manguel Aalberto (1996), *A history of reading*. New York: Penguin Books.

Moretti Franco (2013), *Distant reading*. London; New York: Verso.

Ong Walter J. (1992), *Oralność i piśmienność. Słowo poddane technologii*. Przeł. i wstępem opatrzył Józef Japola. Lublin: Redakcja Wydawnictw Katolickiego Uniwersytetu Lubelskiego.

Paryzek Piotr (2008), Comparison of selected methods for the retrieval of neologisms. *Investigationes Linguisticae* XVI, 163–181.
Patejuk Agnieszka, Przepiórkowski Adam (2015), Parallel development of linguistic resources: towards a structure bank of Polish. *Prace Filologiczne* LXV, 255–270.
Pawłowski Adam (1998), *Séries temporelles en linguistique. Avec application à l'attribution de textes: Romain Gary et Émile Ajar*. Paris, Genève: Champion-Slatkine.
Pawłowski Adam (2001), *Metody kwantytatywne w sekwencyjnej analizie tekstu*. Warszawa: Uniwersytet Warszawski, Katedra Lingwistyki Formalnej.
Pawłowski Adam (2005), *Modelling of the sequential structures in text*. In: Reinhard Köhler, Gabriel Altmann, Rajmund Piotrowski (eds.), *Quantitative Linguistik / Quantitative Linguistics. Ein Internationales Handbuch / An International Handbook*. Berlin, New York: Walter de Gruyter, 738–750.
Pawłowski Adam (2006), Chronological analysis of textual data from the 'Wrocław Corpus of Polish'. *Poznań Studies in Contemporary Linguistics* (PSiCL) 41, 9–29.
Pawłowski Adam, Eder Maciej (2001), Quantity or stress? Sequential analysis of Latin prosody. *Journal of Quantitative Linguistics* 8(1), 81–97.
Pawłowski Adam, Krajewski Marek, Eder Maciej (2010), Time series modelling in the analysis of Homeric verse. *Eos* 47(1), 79–100.
Rokoszowa Jolanta (1989), *Czas a język. O asymetrii zjawisk językowych*. Kraków: Nakładem Uniwersytetu Jagiellońskiego.
Sawicka Grażyna (2006), *Co czas „robi" z językiem?* W: Anna Dąbrowska, A. Nowakowska (red.), *Czas, język, kultura. Język a kultura* 19, 11–32.
Siciński Andrzej (1975), Uwagi o sprawie relatywności czasu. *Człowiek i Światopogląd* 3 (116), 89–102.
Sowa Jan (2008), *Ciesz się, późny wnuku! Kolonializm, globalizacja i demokracja radykalna*. Kraków: Korporacja Ha!art.
Ścibor Eugeniusz (1999), *Wybrane zagadnienia teorii języków informacyjnych*. Olsztyn: Wyższa Szkoła Pedagogiczna.
Vandendorpe Christian (2008), *Od papirusu do hipertekstu: esej o przemianach tekstu i lektury*. Przekł. Anna Sawisz. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.
Viklund Jon, Borin Lars (2016), *How Can Big Data Help Us Study Rhetorical History?* In: De Smedt Koenraad (red.), *Clarin 2015*. Selected Papers from the CLARIN Annual Conference 2015 October 14–16, 2015 Wrocław, Poland, 79–93.
Wierzchoń Piotr (2005), Automatyczne metody ekscerpcji neologizmów, czyli językoznawstwo faktograficzne. *Scripta Neophilologica Posnaniensia* 7, 221–240.

CHRONOLOGICAL CORPORA AND LEXICAL TIME SERIES AS TOOLS FOR DETECTING KEYWORDS AND NEOSEMANTISMS. ON CONCEPTUALIZATION OF TIME IN CHRONOLOGICAL CORPORA

Summary

The article discusses the definition of chronological corpora and lexical series generated thanks to them; moreover, it describes various conceptualizations of categories of time in corpus studies. Chronological corpora are defined in opposition to general and historical corpora. Time is presented in structural perspective (continuity, circularity), anthropological perspective (astronomical, cultural, political, civilizational time) and methodological perspective (models of time series analysis). Taking into account specificity of lexical and statistical data, a model of anomaly (catastrophe) expressed in sudden increase or decrease

of lexeme frequency has been added to used models of sequential phenomena (linear trend, periodic oscillations and random variations). Basing on the analysis of lexical series, the method of automatic identification of potential keywords (they follow a model of anomaly) and neosemantisms (they follow a seriously disturbed model of a steady trend).

Trans. Izabela Ślusarek